Regression on the trees data with R

```
> trees
   Girth Height Volume
1    8.3     70   10.3
2    8.6     65   10.3
3    8.8     63   10.2
4   10.5     72   16.4
5   10.7     81   18.8
6   10.8     83   19.7
7   11.0     66   15.6
8   11.0     75   18.2
9   11.1     80   22.6
10  11.2     75   19.9
11  11.3     79   24.2
12  11.4     76   21.0
13  11.4     76   21.4
14  11.7     69   21.3
15  12.0     75   19.1
16  12.9     74   22.2
17  12.9     85   33.8
18  13.3     86   27.4
19  13.7     71   25.7
20  13.8     64   24.9
21  14.0     78   34.5
22  14.2     80   31.7
23  14.5     74   36.3
24  16.0     72   38.3
25  16.3     77   42.6
26  17.3     81   55.4
27  17.5     82   55.7
28  17.9     80   58.3
29  18.0     80   51.5
30  18.0     80   51.0
31  20.6     87   77.0

>
> help(trees)
>
```

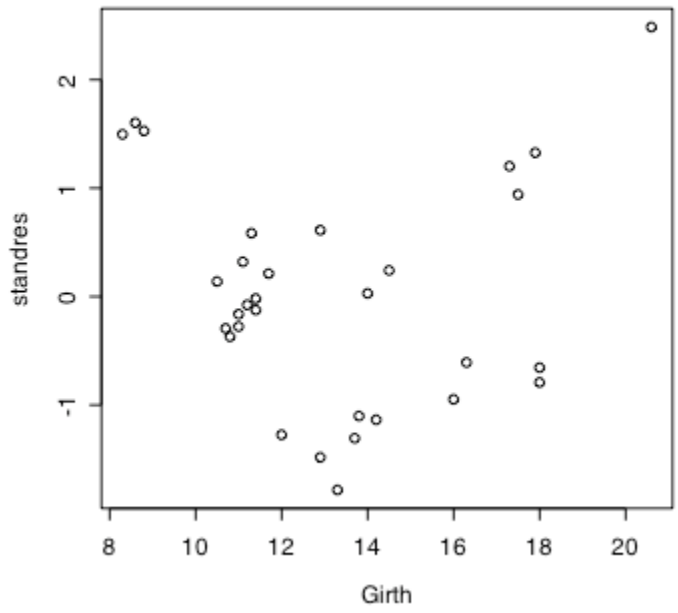
A data frame with 31 observations on 3 variables.

[,1] Girth	numeric	Tree diameter in inches
[,2] Height	numeric	Height in ft
[,3] Volume	numeric	Volume of timber in cubic ft

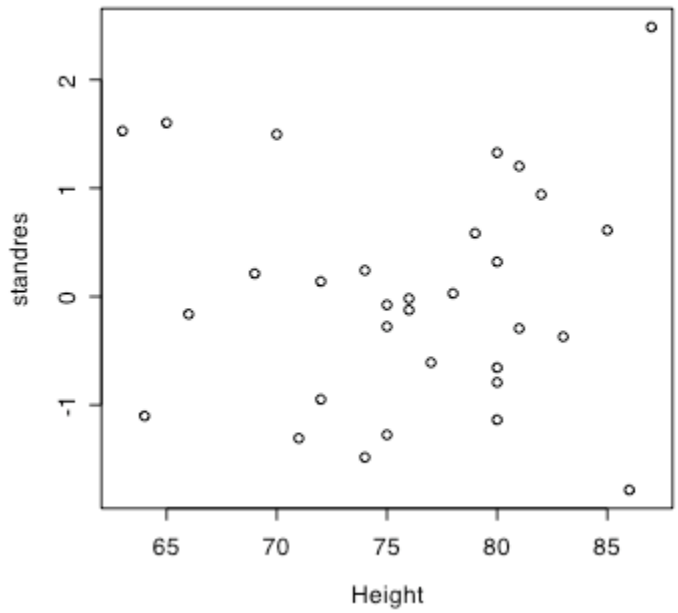
```
> ls()
character(0)
> Girth
Error: object "Girth" not found
> attach(trees)
> ls()
character(0)
> Girth
 [1]  8.3  8.6  8.8 10.5 10.7 10.8 11.0 11.0 11.1 11.2 11.3 11.4 11.4 11.7 12.0
[16] 12.9 12.9 13.3 13.7 13.8 14.0 14.2 14.5 16.0 16.3 17.3 17.5 17.9 18.0 18.0
[31] 20.6
```

```
> treemod1 = lm(Volume ~ Girth + Height)
> # Alternative: treemod1 = lm(Volume ~ Girth + Height, data=trees)
>
> summary(treemod1)

Call:
lm(formula = Volume ~ Girth + Height)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4065 -2.6493 -0.2876  2.2003  8.4847

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
Girth         4.7082     0.2643  17.816  < 2e-16 ***
Height        0.3393     0.1302   2.607   0.0145 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-Squared: 0.948,    Adjusted R-squared: 0.9442
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16

> # Those are unstandardized residuals
>
> standres = rstandard(treemod1) # Actually, Studentized
> studres = rstudent(treemod1) # Actually,  Studentized deleted
> cbind(treemod1$residuals,standres,studres)
                  standres     studres
1   5.46234035  1.49649007  1.53206937
2   5.74614837  1.60294618  1.65166828
3   5.38301873  1.52845547  1.56773982
4   0.52588477  0.13967002  0.13720104
5  -1.06900844 -0.29367511 -0.28882839
6  -1.31832696 -0.36961632 -0.36384474
7  -0.59268807 -0.16228164 -0.15943240
8  -1.04594918 -0.27666283 -0.27204961
9   1.18697860  0.32089637  0.31569502
10 -0.28758128 -0.07592750 -0.07456700
11  2.18459773  0.58477425  0.57777591
12 -0.46846462 -0.12369228 -0.12149660
13 -0.06846462 -0.01807723 -0.01775159
14  0.79384587  0.21237488  0.20871616
15 -4.85410969 -1.27469222 -1.28970287
16 -5.65220290 -1.48274728 -1.51679495
17  2.21603352  0.61250123  0.60553457
18 -6.40648192 -1.78323847 -1.85990126
19 -4.90097760 -1.30685072 -1.32432594
20 -3.79703501 -1.10137319 -1.10574384
21  0.11181561  0.02933487  0.02880672
22 -4.30831896 -1.13596377 -1.14212275
23  0.91474029  0.24176173  0.23765348
24 -3.46899800 -0.94802613 -0.94625363
25 -2.27770232 -0.60821465 -0.60123981
26  4.45713224  1.20259894  1.21266187
27  3.47624891  0.94188356  0.93992125
28  4.87148717  1.32756957  1.34672046
29 -2.39932888 -0.65511219 -0.64829498
30 -2.89932888 -0.79163207 -0.78621538
31  8.48469518  2.48614353  2.76560250
```

```
> # Plot standardized residuals vs vars in model
> plot(Girth,standres)
```



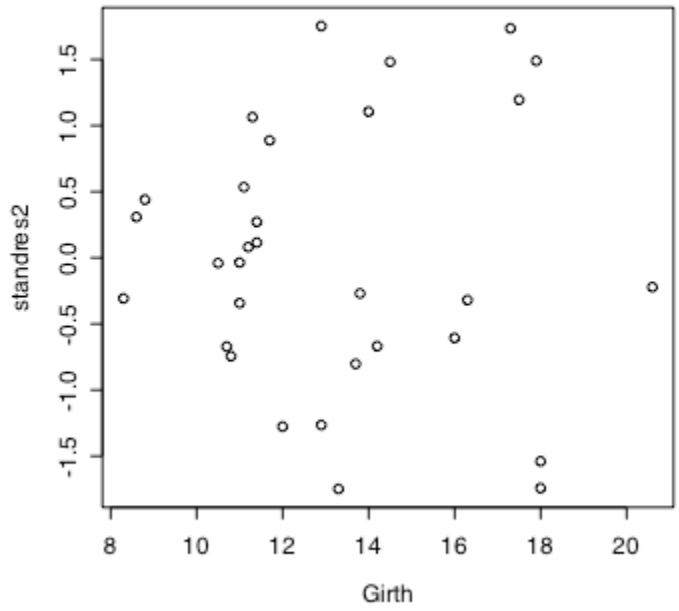

```
> plot(Height,standres)
```



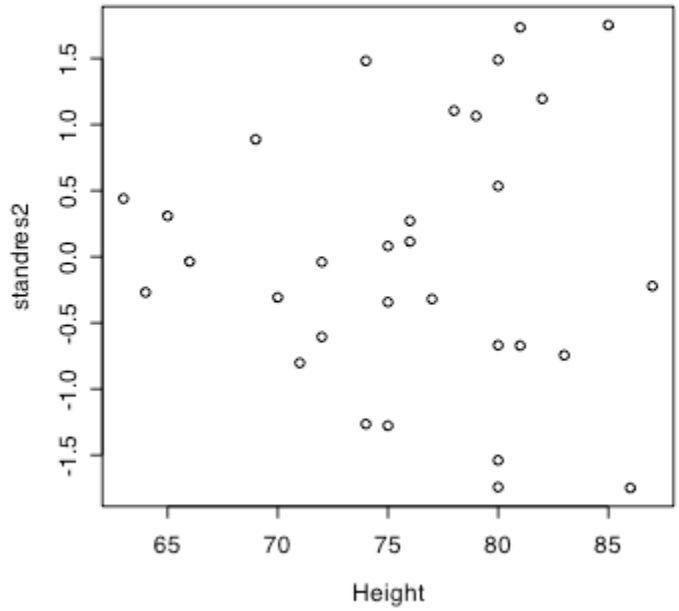

```
> # First plot has a clear U-shape, and it makes sense
```

```
> Girthsq = Girth^2
> treemod2 = lm(Volume ~ Girth + Girthsq + Height)
> summary(treemod2)

Call:
lm(formula = Volume ~ Girth + Girthsq + Height)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2928 -1.6693 -0.1018  1.7851  4.3489

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.92041   10.07911  -0.984 0.333729
Girth       -2.88508    1.30985  -2.203 0.036343 *
Girthsq      0.26862    0.04590   5.852 3.13e-06 ***
Height       0.37639    0.08823   4.266 0.000218 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.625 on 27 degrees of freedom
Multiple R-Squared: 0.9771,   Adjusted R-squared: 0.9745
F-statistic: 383.2 on 3 and 27 DF,  p-value: < 2.2e-16

> # Another way to test H0: beta2=0
> anova(treemod1,treemod2)
Analysis of Variance Table

Model 1: Volume ~ Girth + Height
Model 2: Volume ~ Girth + Girthsq + Height
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     28 421.92
2     27 186.01  1    235.91 34.243 3.13e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> 5.852^2 # F = t^2
[1] 34.24590
> summary(treemod2)$coefficients[3,3]^2
[1] 34.24275

> anova(treemod2)
Analysis of Variance Table

Response: Volume
          Df Sum Sq Mean Sq  F value    Pr(>F)
Girth      1 7581.8  7581.8 1100.511 < 2.2e-16 ***
Girthsq    1  212.9   212.9   30.906 6.807e-06 ***
Height     1  125.4   125.4   18.198 0.0002183 ***
Residuals 27  186.0     6.9
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> # Numerator SS are sequential; denominator from the full model
> 212.9/6.9
[1] 30.85507
```

```
> standres2 = rstandard(treemod2)
> plot(Girth,standres2)
```

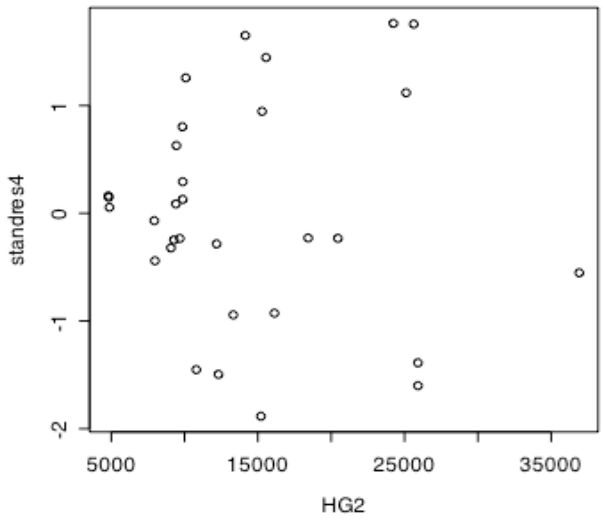



```
> plot(Height,standres2)
```

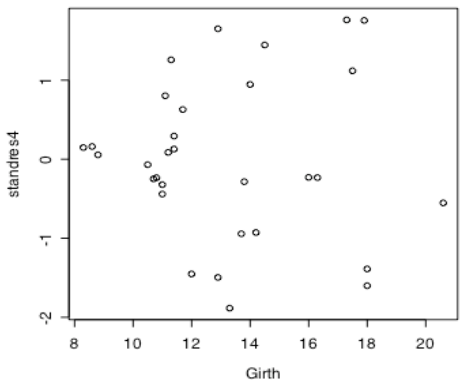


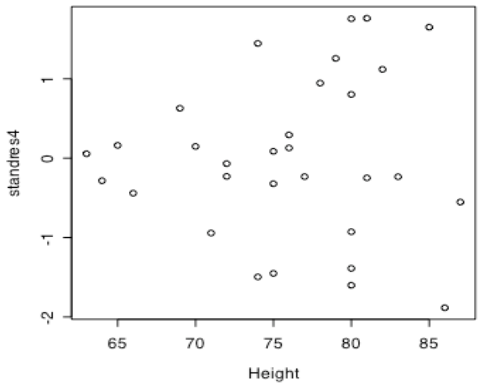
```
> # The interaction of height by girth^2 has a physical meaning: v = pi r-sq h
> HG2 = Height*Girthsq
> treemod3 = lm(Volume ~ Girth + Girthsq + Height + HG2)
> summary(treemod3)

Call:
lm(formula = Volume ~ Girth + Girthsq + Height + HG2)

Residuals:
    Min      1Q  Median      3Q     Max
-4.8268 -1.1152 -0.1531  1.7353  4.2208

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.522657  12.474662  -0.202    0.841
Girth       -0.494555   2.713189  -0.182    0.857
Girthsq      0.036459   0.235294   0.155    0.878
Height       0.075559   0.311769   0.242    0.810
HG2          0.001866   0.001854   1.006    0.324

Residual standard error: 2.624 on 26 degrees of freedom
Multiple R-Squared: 0.9779,   Adjusted R-squared: 0.9745
F-statistic: 287.8 on 4 and 26 DF,  p-value: < 2.2e-16

> # No variable is significant controlling for all the others

> treemod4 = lm(Volume ~ HG2); summary(treemod4)

Call:
lm(formula = Volume ~ HG2)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6195 -1.1002 -0.1656  1.7451  4.1976

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.977e-01  9.636e-01  -0.309     0.76
HG2          2.124e-03  5.949e-05  35.711   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.493 on 29 degrees of freedom
Multiple R-Squared: 0.9778,   Adjusted R-squared: 0.977
F-statistic:  1275 on 1 and 29 DF,  p-value: < 2.2e-16

> anova(treemod4,treemod3)
Analysis of Variance Table

Model 1: Volume ~ HG2
Model 2: Volume ~ Girth + Girthsq + Height + HG2
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     29 180.236
2     26 179.042  3     1.193 0.0578 0.9814
> # I like Model 4
```

```
> # Plot resid vs var in model
> standres4 = rstandard(treemod4)
> plot(HG2,standres4)
```

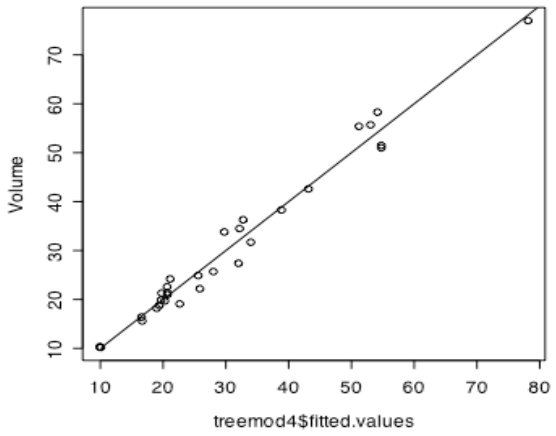



```
> # Plot resid vs vars NOT in model
> plot(Girth,standres4)
> plot(Height,standres4)
```

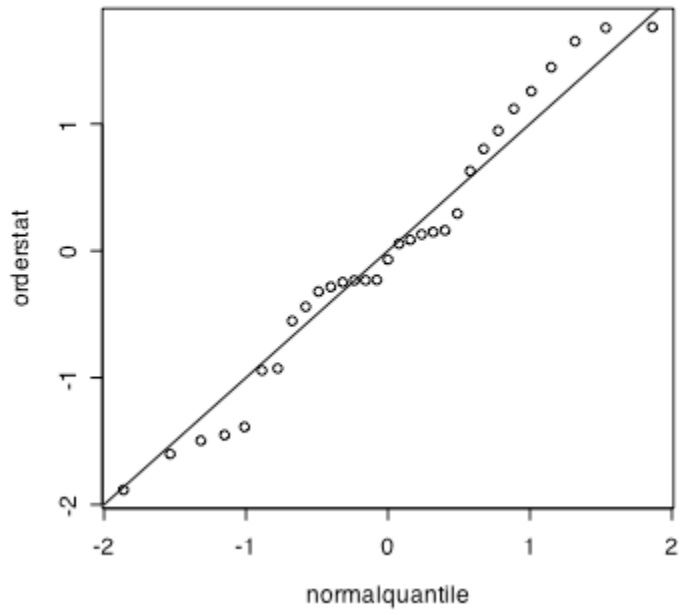




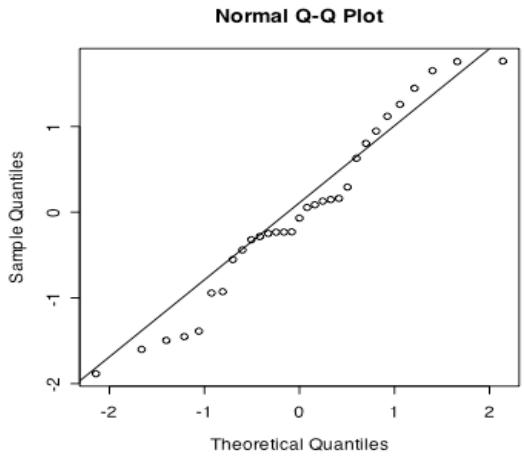
```
> # Plot Y vs Y-hat
> plot(treemod4$fitted.values,Volume)
> lines(c(10,80),c(10,80))
> cor(treemod4$fitted.values,Volume)^2 # Equals R^2
[1] 0.9777654
```



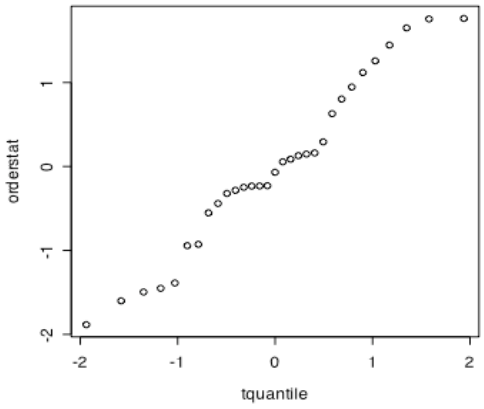

```
> # Normal QQ plot of (standardized) residuals
> orderstat = sort(standres4)
> n = nrow(trees)
> quants = (1:n)/(n+1)
> normalquantile = qnorm(quants)
> plot(normalquantile,orderstat)
> lines(c(-2,2),c(-2,2))
```



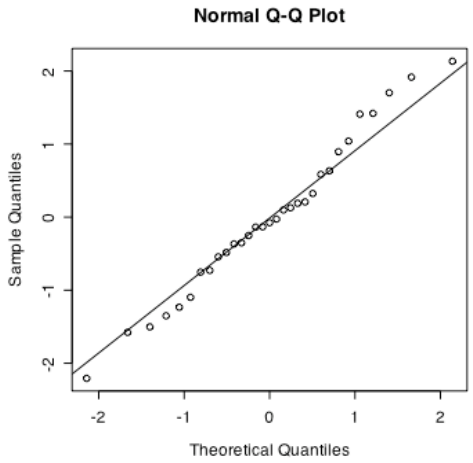
```
> # That does not look very good. An automated way ...
> qqnorm(standres4)
> qqline(standres4)
> # qqline goes through 1st and 3d quantiles
```




```
> # t quantiles?
> tquantile = qt(quants,n-2)
> plot(tquantile,orderstat)
```




```
> z = rnorm(31)
> qqline(z)
> qqnorm(z)
> qqline(z)
```



```
> help(predict.lm)

predict(object, newdata, se.fit = FALSE, scale = NULL, df = Inf,
        interval = c("none", "confidence", "prediction"),
        level = 0.95, type = c("response", "terms"),
        terms = NULL, na.action = na.pass, pred.var = res.var/weights,
        weights = 1, ...)

> # Predict for a tree 75 ft tall, 10 in around
> newtree = data.frame(HG2 = 75*10^2)
> predict(treemod4,newtree)
[1] 15.63513

> predict(treemod4,newtree,interval="prediction")
          fit      lwr      upr
[1,] 15.63513 10.38833 20.88193

> # Is this what I think it is?
> treemod4$coefficients
 (Intercept)          HG2
-0.297679437  0.002124374
> treemod4$coefficients[1] + treemod4$coefficients[2]*7500
(Intercept)
   15.63513

# Now reproduce the interval
```

$$1-\alpha = Pr\left\{\mathbf{x}'_{n+1}\widehat{\boldsymbol{\beta}} - t_{\alpha/2}\, s\{d_{n+1}\} < Y_{n+1} < \mathbf{x}'_{n+1}\widehat{\boldsymbol{\beta}} + t_{\alpha/2}\, s\{d_{n+1}\}\right\}$$

Need $s\{d_{n+1}\}$. Formula for deleted residual is inconvenient.

$$\begin{aligned} d_{n+1} &= Y_{n+1} - Y_{n+1(n+1)} \\ V(d_{n+1}) &= V(Y_{n+1}) + V(\mathbf{x}'_{n+1}\widehat{\boldsymbol{\beta}}) \\ &= \sigma^2 + \mathbf{x}'_{n+1} V(\widehat{\boldsymbol{\beta}}) \mathbf{x}_{n+1} \\ &= \sigma^2 + \mathbf{x}'_{n+1} \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_{n+1} \\ &= \sigma^2 \left(1 + \mathbf{x}'_{n+1} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_{n+1}\right) \end{aligned}$$

Need $(\mathbf{X}'\mathbf{X})^{-1}$. Estimate σ^2 with sqrt(MSE)

$$1-\alpha = Pr\left\{\mathbf{x}'_{n+1}\widehat{\boldsymbol{\beta}} - t_{\alpha/2}\, s\{d_{n+1}\} < Y_{n+1} < \mathbf{x}'_{n+1}\widehat{\boldsymbol{\beta}} + t_{\alpha/2}\, s\{d_{n+1}\}\right\}$$

$$V(d_{n+1}) \quad = \quad \sigma^2\left(1 + \mathbf{x}'_{n+1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_{n+1}\right)$$

```
> treemod4 = lm(Volume~HG2,x=T) # Include X matrix in model object
> X = treemod4$x; xpxinv = solve(t(X)%*%X)
> mse = anova(treemod4)[2,3]; mse
[1] 6.215032
> newx = c(1,75*10^2)
> pred = sum(newx*treemod4$coefficients); pred # Should be 15.63513
[1] 15.63513
> dim(newx) = c(2,1); newx
     [,1]
[1,]    1
[2,] 7500
> sepred = sqrt( mse * (1 + t(newx)%*%xpxinv%*%newx)); sepred
         [,1]
[1,] 2.565385
> tcrit = qt(0.975,29); tcrit
[1] 2.045230
> # Upper prediction limit
> pred+sepred*tcrit
         [,1]
[1,] 20.88193
> predict(treemod4, newtree, interval="prediction")
          fit      lwr      upr
[1,] 15.63513 10.38833 20.88193
> # That's it!
```